Robert_S3_L08

Thu, 2/17 4:01PM 11:22

SUMMARY KEYWORDS

correlation, correlation coefficient, variables, simple linear regression, outliers, relationship, correlation matrix, data, observations, statistical technique, smoking, lung cancer, negative, causation, series, cancer, tobacco companies, important, correlated, unit

SPEAKERS

Robert McKeown

R Robert McKeown 00:04

Hello, and welcome back. In this video, we're going to be looking at the properties of correlation. It's important for us to review correlation because it's important and shows up very often in the social sciences, especially in this day and age, when we have so much data big data available, we're always looking to apply statistical techniques to that data. And it's important you understand the advantages to doing that, and the disadvantages, the strengths and weaknesses of the correlation coefficient. But also, it's going to lead us into simple linear regression, which is a technique for making predictions about say, a Y valuable- value based on an X value. And we're going to talk about that in just a few moments. But first, let's review correlation, because it really makes the foundation for simple linear regression. Taking a look at the slide in front of you, we can see that the first bullet point for the properties of correlation is that the correlation coefficient is always between negative one and one. If you see a correlation coefficient that is greater than one or less than negative one, then you know that you have a problem. We often say that if the correlation coefficient is negative one, it's perfectly negatively correlated. And if the correlation coefficient is plus one, it's perfectly positively correlated, or just perfectly correlated. So there's the range of values that correlation can take on. The correlation is a statistic between two variables. And if we have the correlation of X with Y, we know that the correlation of Y with X is going to be the same. And so whether we're talking whichever variable we put first, and whichever second, it's kind of irrelevant, because it's reciprocal to the correlation of X, and Y is equal to the correlation of Y with X. Now if we have more than two series values, or we have more than two variables, variable being a series data, then we need to use the correlation matrix, which we saw in an earlier video, the correlation matrix will give us the correlation between each to each pair of variables. Remember, we saw something like if we had X, Y, Z, X, Y, Z, then we might have we're gonna have one on the diagonal axes, we might have say, point five here, if Y and X have a correlation a point five, then x and y is also going to have a correlation of point five. And if Z and X have a correlation of point two, then x and z is also going to have a correlation of point two. If Y and Z have a correlation of point three, then Z and Y is going to have a correlation of point three. There's the correlation matrix. I've labeled it for you. And I put some lines into the matrix to help you see the how, which number is matched to which variable. So the X, Y and Z are variables here. This is how we would show correlation between more than two variables. The sine of the correlation

coefficient determines the type of relationship so if we have a correlation equal to negative 0.57. That's going to be a negative relationship that's gonna have as a negative correlation, or the two series are negatively correlated. If the correlation coefficient is positive, it's going to be a positive relationship and whether the values the number Whether these numbers in here are greater or lesser than each other doesn't matter, it's the sign that tells us the relationship. Correlation is not a unit of measure. That's in the sense that correlation is not representative of Celsius. It's not a percentage, either. It is a number between negative one and plus one. And that range. This means that whether the underlying units are represented in say, Celsius, or Fahrenheit, it's not going to affect the correlation coefficient. So it does not change the correlation coefficient. If you transform the underlying series from one unit of measure to another, or say you multiply both the the series or you multiply a series by 100. That is not going to change the correlation coefficient as long as you are consistent in your transformation. And the punch line here is always present correlation without units, and not as a percentage. Sometimes people do present it as a percentage, but it's not really appropriate. Correlation should be between negative one and one and just state it as a decimal. I'm sure, as is shown up in previous videos, but a very common refrain in all disciplines that use statistics is do not confuse correlation with causation. Notice that the correlation of X and Y is also the correlation of Y with X. So suppose that we're looking at smoking and cancer. And the correlation between smoking and cancer is say, but let's just hypothesize that it's 0.8, specifically, lung cancer. That alone doesn't tell you that smoking causes cancer. That just means that if you have cancer or lung cancer, specifically, you are likely to have or tend to be a smoker. And in fact, there is in the 20th century, there were court cases, trying to prove that smoking caused cancer. And statisticians for tobacco companies that were defending the tobacco companies often argued that or occasionally argued that people with lung cancer like to smoke because it relieve their symptoms, which to you and I, and in the 21st century, that sounds like a ridiculous argument, which it is, but just because they they had correlation between the two, they actually couldn't show there's causation. Later, they did use many studies and random control trials to show that, indeed, smoking causes lung cancer, but based on the correlation alone, cannot be certain. As we saw in an earlier video, correlation can be misleading. If the two quantitative variables have a nonlinear relationship. In this situation, you don't want to use our correlation coefficient, you would want to use some more sophisticated statistical technique, possibly multiple regression. Correlation is sensitive to outliers. So if we have an outliers, unusual observations, or observations that are far away from the mean, then that can make a very strong relationship appear weak, or it could make a weak relationship appear strong. So when we're looking at correlation coefficients, we always want to be aware of the influence of outliers. Sometimes, you can even see that by removing one or two observations, the correlation coefficient will change drastically. This is problematic because there is no way to deal with outliers every single time that we think we identify why we can't even Come up with a strict definition of an outlier. One that we know for sure is an outlier. That works in all situations. Often as a practitioner, when you're doing research, you might want to look at those specific data points, and figure out if there's a measurement error there, or trying to understand why these observations are so different than the other observations within your series and data set. Lastly, we will look and we'll continue to study the correlation determination of correlation determination measures the variability in one variable that is explained by the variability in another. And since it's a correlation squared, the it two has nothing to do with causation in the form that we're seeing it here. So make sure that you do not confuse correlation with causation. We'll see the R squared measure again. Now that we look at linear regression